Remarks

Claims 1-20 are pending. Claims 1-20 are rejected. Claim 13 is canceled. All rejections are traversed. Claim 21 is new. No new subject matter is introduced.

Claims 1-5, 8, 11-14, 17, 19 and 20 are rejected under 35 U.S.C. 102(b) as being anticipated by Leonardi et al., "Semantic indexing of multimedia documents" (Leonardi).

Leonardi divides an input video stream into an audio and a video component, see page 46: "First, we divide the input stream into the two main components - audio and video." Each stream is then classified by a HMM, see page 47: "We then classify each sequence of feature vectors extracted from the two streams by an HMM used in an innovative approach."

Thus, in Leonardi there are ***two distinct and separate*** HMMs: an audio HMM for the audio features and a visual HMM for the visual features.

Leonardi Audio HMM

The audio HMM is used to segment the video into audio classes, see page 47: "For audio classification we considered four classes - namely music, silence, speech, and background noise... By using the HMM model, the algorithm can estimate the optimal sequence of hidden states representing the different audio classes associated to the different temporal frames.

Finally, consecutive frames marked by the same class define the various audio segments."

Leonardi Visual HMM

The visual HMM is used to segment the video into shots, see page 47: "In the video analysis, the system segments the video signal into elementary units, which form individual video shots. With this aim, we train a two-state HMM classifier, in which it associates S1 as a detected shot transition and S0 as a non-detected shot transition… The adopted approach to perform scene identification requires us to define four different types of scenes…"

It is clear that the operations on the two HMMs in Leonardi are independent of each other. That is, the audio classification only uses the audio HMM, and the scene classification only uses the visual HMM. Neither HMM depends on the other.

In contrast, the claimed invention fuses probabilistically the audio labels and visual labels into a single discrete-observation coupled hidden Markov model (DCHMM) to detect highlights in the video, see Figure 4. Thus, the invention uses a **single** DCHMM, instead of two separate HMMs as in Leonardi. As advantage, the fusing of the audio and visual features into a single model makes the detection of highlights jointly dependent on **both the audio and visual features**. That is, the invention can leverage the interaction of the audio and visual signals. An example of this advantage is given in the present application at paragraph [033]: "For instance, in golf,

high (visual) motion caused by a good shot is often followed by (audio) applause." This indicates a transition of a visual state (swing) to a following audio state (applause), *i.e.*, a highlight.

This joint analysis can never be achieved by Leonardi.

DCHMMs have been known for over a decade. For a general background description of DCHMMs see, Matthew Brand, "Coupled hidden Markov models for modeling interactive processes," Tech. Rep. 405, MIT Media Lab, 1997, incorporated herein by reference. DCHMMs have never been used for detecting highlights in videos. There is nothing in Leonardi that even remotely resembles a DCHMM as known in the art.

Moreover, Leonardi does not perform highlight detection. Leonardi only classifies scenes: "For each kind of scene, we've calculated four different performance indices: completeness, purity, $\text{completeness}_{\text{cover}}$ and $\text{purity}_{\text{cover}}$. We defined the first two scene classes (completeness and purity) as in the case of the audio and video classification simulations, while we introduced the latter two ($\text{completeness}_{\text{cover}}$ and $\text{purity}_{\text{cover}}$) to consider situations when the system recognizes some shots correctly and improperly recognizes others." Classifying scenes in a video does not detect highlights in the video as claimed.

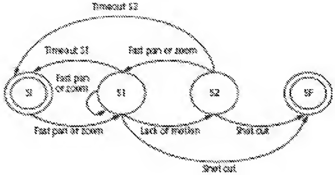
With respect to claim 2, Leonardi does not describe a discrete-observation coupled hidden Markov model including the audio features, the visual features, states of the audio features and states of the visual features.

With respect to claims 3, 4, 8 and 11, Leonardi does not fuse silent features into a DCHMM with visual features.

With respect to claim 5, Leonardi only says that his method is applied to MPEG-7 content, see page 50: "Considering the bottom-up approach, we analyzed several samples from the MPEG-7 content set using the proposed classification schemes."

These descriptors are defined in "International Organisation for Standardisation ISO/IEC/ JTC1/SC29/WG11 CODING OF MOVING PICTURES AND AUDIO INFORMATION ISO/IEC JTC1/SC29/WG11/ N7709 Nice, October 2005. There is nothing in Leonardi that indicates that MPEG7 audio descriptors are used for his classification. Leonardi only uses low level features, see page 46: "This step segments the audio stream into clips, and extracts a feature vector from the low level acoustic properties of each clip (such as the Mel-Cepstrum coefficients, zero crossing rate, and so on)." With all due respect, the Examiner's statement is not supported by Leonardi. Applicants cannot find the use of MPEG-7 audio descriptors in Leonardi.

With respect to claim 12, Leonardi does not disclose the visual labels selected from the group consisting of close shot and replay.



With respect to claim 14, the HMM described at page 47, lines 31-41, only pertains to the audio HMM, see above: "These define the observations produced by an ergodic HMM. By using the HMM model, the algorithm can estimate the optimal sequence of hidden states representing the different audio classes associated to the different temporal frames. Finally, consecutive frames marked by the same class define the various audio segments." There is no indication in Leonardi of a DCHMM that includes both audio and visual features.

4. Claims 6, and 7 are rejected under 35 U.S.C. 103(a) as being unpatentable over Leonardi et al., (Semantic Indexing if Multimedia Documents, April –June 2002), in view Rui et al., (Automatically Extracting Highlights for TV Baseball Programs, Eighth ACM International Conference on Multimedia, pp.105 – 115, 2000)

With respect to claim 6, Rui only describes Gaussian fitting, see pages 113-114: "In Section 4.3, we discussed three approaches to excited speech classification: Gaussian fitting (*GAU*), K nearest neighbors (*KNN*), and

support vector machines (*SVM*)." There are numerous ways that Gaussians can be fitted, but there is no indication in Rui that the fitting uses Gaussian mixture models as claimed.

5. Claims 9, 10, and 15 are rejected under 35 U.S.C. 103(a) as being unpatentable over Leonardi et al., (Semantic Indexing if Multimedia Documents, April –June 2002), in view Wang et al., (Integration of Multimodal Features For Video Scene Classification based on HMM, 1/99).

With respect to claims 9 and 10, Wang does not describe visual features which include dominant color and motion vector and quantized motion activity in a DCHMM. Wang does not teach a DCHMM as claimed.

6. Claims 16, and 18 are rejected under 35 U.S.C. 103(a) as being unpatentable over Leonardi et al., (Semantic Indexing if Multimedia Documents, April –June 2002), in view of Rui et al., (US PAP 2003/0103647).

With respect to claim 16, neither Leonardi nor Rui describe training a discrete-observation coupled hidden Markov model with hand labeled videos.

With respect to claim 18, Leonardi does no teach detecting highlights nor DCHMMS. Rui does determine likelihoods for highlights, and thresholding the highlights. Rui only describes detecting and tracking of faces.

The relevance of the Examiner's following statement with respect to what is claimed is not understood.

> Therefore it would have been obvious to one of ordinary skill in the art at the time the invention was made to threshold candidate face regions as taught by Rui et al., in Leonardi et al., because that would help determine particular scenes by rejecting non relevant scenes.

There is nothing in Leonardi or what is claimed that pertains to rejecting non-relevant scenes.

It is believed that this application is now in condition for allowance. A notice to this effect is respectfully requested. Should further questions arise concerning this application, the Examiner is invited to call Applicants' attorney at the number listed below. Please charge any shortage in fees due in connection with the filing of this paper to Deposit Account 50-0749.

Respectfully submitted,
Mitsubishi Electric Research Laboratories, Inc.

By
/Dirk Brinkman/

Dirk Brinkman
Attorney for the Assignee
Reg. No. 35,460

201 Broadway, 8th Floor
Cambridge, MA 02139
Telephone: (617) 621-7517
Customer No. 022199